Validation of three postoperative risk prediction models for ICU mortality after cardiac surgery

Author's names and details removed as per instructions

This work has not been presented elsewhere. There is no conflict of interests.

Keywords: Postoperative Care, Statistics, Outcomes (includes mortality, morbidity)

Word count:4,385

Introduction

Preoperative cardiac surgery risk prediction models such as the EuroSCORE models [1-3] have been widely adopted and extensively studied. The primary purpose of such models is to allow the risk of mortality to be estimated prior to intervention. The risk estimates can be used to inform clinical decision making when considering intervention and to risk-adjust surgical outcome data on an 'intention to treat' basis.

However, once an intervention has taken place preoperative models become less useful because intraoperative and postoperative events that may affect risk are not taken into account. Following intervention, models which include risk factors from the intra- and postoperative periods may be more useful for estimating risk and could aid postoperative clinical decision making and allow benchmarking of Cardiac Intensive Care Unit (CICU) performance.

A number of potentially useful models which analyse postoperative data have been developed. Some models designed for use in general intensive care unit (ICU) patients also accurately predict mortality after cardiac surgery,[4-14] with the Sequential Organ Failure Assessment (SOFA) score generally demonstrating the best performance.[7, 15] The Cardiac Surgery Risk Score (CASUS) and its derivatives are examples of models designed specifically for use following cardiac surgery. The CASUS model has been validated in Germany using data from multiple institutions [4,6,11,12,16] and in a study of 150 patients in Greece.[17] The derivative Logistic Cardiac Surgery Risk Score (logCASUS)[18] and Rapid Clinical Evaluation (RACE)[19] models which calculate ICU mortality risk have not been externally validated.

Despite a number of models being available, few are utilised in clinical practice. This lack of adoption may be due to the absence of comparative external validation studies in contemporary cohorts. The objective of this study was therefore to validate the

logCASUS, RACE and SOFA scores for the prediction of ICU mortality in cardiac surgery patients. The performance of serial daily scores for each model was also assessed.

Patients and Methods

Prospectively collected data for consecutive adult patients admitted to the CICU at our institution following cardiac surgery between 1st January 2013 and 31st May 2015 were analysed. Our institution is a tertiary adult cardiac surgery centre and our case-mix includes patients undergoing cardiac transplantation and mechanical circulatory support. As in the original studies which described the logCASUS[18] and RACE[19] scores, only data from each patient's first CICU admission after cardiac surgery were included. Patients whose CICU admissions were too short to allow calculation of risk using the models were excluded. The primary outcome for the study was ICU mortality.

DATA COLLECTION, VALIDATION AND CLEANING

Preoperative patient characteristics and postoperative outcome data were collected from the clinical governance database which is compiled by clinicians and validated by database managers. Postoperative data from the patients' CICU admissions were obtained from the electronic patient record. Results of blood analyses were obtained from the pathology laboratory database and data concerning postoperative cerebrovascular accidents were obtained from the radiology database. As described in the original studies [15,18,19] the most abnormal value for each variable recorded each day was entered into the models. Data from all four sources were collated and cleaned using reproducible algorithms in R Studio (R Foundation for Statistical Computing)[20].

All data were entered into the Vascular Governance North West database and managed according to the protocol and ethical approvals governing this database. As

data were pseudonymised prior to analysis the Research Ethics Committee concluded that ethical approval for these analyses was not necessary.

MISSING DATA

Where a variable was not measured on a given day, the patient's most recent postoperative value was used to calculate the risk score. Except for bilirubin, substituted blood test data was required in <3% of risk score calculations. Previous postoperative bilirubin concentrations were not available for 8.3% of daily calculations and therefore missing values were substituted using the nearest subsequent value for that patient. Where the above substitutions were not possible due to a complete absence of data for a given patient, the median value for that parameter in all patients was imputed. Bilirubin was imputed in this way for 6.2% of patients but other variables were only imputed for 0.1% of patients.

For the logCASUS score, calculation of the pressure adjusted heart rate (PAR) (which combines information from heart rate, central venous pressure (CVP) and mean arterial pressure) was not possible for 7% of score calculations. This was most commonly because the central venous catheter had been removed before CICU discharge. To address missing CVP data, a logistic regression model was developed using data from patients for whom data was complete. This model was then used to calculate a modelled PAR. There were no missing outcome data.

STATISTICAL ANALYSES

Central tendency of variables is described using mean and standard deviation where the distribution was parametric and median and interquartile range where the distribution was non-parametric.

The logCASUS, RACE and SOFA scores were calculated for each patient on a daily basis for postoperative days one to seven. The discrimination of all scores for the

prediction of ICU mortality was assessed using the area under the receiver operating characteristic curve (AUC). De Long's method for calculating AUC variance was used for the calculation of AUC 95% confidence intervals.[21] AUC values of ≥0.7 were considered acceptable, ≥0.8 was considered good.

The calibration of logCASUS and RACE ICU mortality estimates was assessed using the ratio of observed outcomes to predicted outcomes (O:E ratio), the Hosmer Lemeshow (HL) test and calibration plots. A high HL χ^2 value with a low p value suggests that there is a significant difference between predicted risk and observed outcomes.[22] The calibration plots illustrate how the mean predicted probability of ICU mortality compares with the observed incidence of ICU mortality for five equally sized groups based on the ranked predicted risks calculated by the model. Calibration of the original SOFA score could not evaluated because it is a non-logistic score.

A sub-group analysis excluding patients who underwent cardiac transplantation or initiation of mechanical circulatory support was also performed. Finally, local recalibration of the models was performed. Data were divided randomly into two equally sized datasets; a training dataset and an evaluation dataset. Each model was fully recalibrated using data for each variable from the training dataset. The calibration and discrimination of each recalibrated model was then tested in the evaluation dataset.

Results

PATIENT CHARACTERISTICS

There were 2284 consecutive patients who met the inclusion criteria. 29 patients were excluded because their admission to CICU was too short to allow calculation of the risk scores. The mean (sd) age of the patients was 65.7 (11.8) and 27.3% were female.

The most common procedure was isolated coronary artery bypass graft surgery (53.3%). Additional patient and operative characteristics are shown in Table 1. The overall ICU mortality rate was 2.0%. The ICU mortality rate in the final validation cohort was 1.8%.

MODEL PERFORMANCE ON THE FIRST POSTOPERATIVE DAY

The variables included in each model are detailed in Table 2. A day-by-day description of the levels of risk predicted by the models is shown in Appendix A. All three models demonstrated good discrimination when calculated on the first postoperative day (Figure 1a). The AUC for the RACE and logCASUS scores were the same at 0.94 (95%CI 0.91-0.97) for both. The AUC for the SOFA score was 0.91 (95%CI 0.86-0.96). The HL tests, together with the comparison of the O:E ratios and calibration plots implied poor calibration of both logistic models (Table 3). As seen in Figure 1b, predictions were least accurate for those patients who had the highest predicted risk. Sub-group analysis demonstrated no significant effect on model performance when patients undergoing cardiac transplantation or initiation of mechanical circulatory support were excluded.

SERIAL SCORES

The daily measures of discrimination and calibration for the models are shown in Table 3. LogCASUS and RACE scores calculated daily up to day seven of the postoperative CICU admission demonstrated good discrimination. The AUC of the SOFA score was generally lower than those of the cardiac surgery-specific scores in the early postoperative period but the difference reduced towards the end of the first postoperative week. Calibration plots, HL test and O:E ratios suggested poor calibration for logCASUS. Calibration for RACE was better but remained suboptimal.

LOCAL RECALIBRATION

The AUC, O:E ratios and HL test results for the recalibrated models' performance in the evaluation dataset are detailed in Table 4. The analyses of recalibrated model performance were limited to the first five days as the training dataset only contained seven patients who died after being on CICU for more than 5 days. Local recalibration of models using the training dataset generally resulted in marginal improvement in discrimination for all scores. The calibration of the recalibrated logCASUS model was adequate on every day. The HL tests for RACE and SOFA showed adequate calibration on every day except day 5. Calibration plots for the original and recalibrated models are shown in Figure 1b. Full details of the recalibrated models are provided in appendices B-D.

Discussion

This study has validated the performance of the logCASUS, RACE and SOFA scores in a cohort of 2255 patients from a tertiary cardiac centre in the UK. The observed ICU mortality (1.8%) is in line with that for all cardiac surgery in the UK [23]. This is despite the cohort including 62 patients who underwent emergency/salvage procedures and 41 who underwent instigation of mechanical circulatory support. In these groups ICU mortality was 21.0% and 24.3% respectively. In the remaining patients the ICU mortality rate was 1.2%.

All models demonstrated good discrimination throughout the first postoperative week with discrimination declining slightly towards the end of the week. Both the logCASUS and RACE scores demonstrated poor calibration in our cohort and significantly over-predicted risk. The poor calibration demonstrated by the RACE and CASUS models may be due to a similar calibration drift effect to that observed with preoperative risk models due to improvements in care and clinical outcomes over time.[24] Assuming

that our findings of poor calibration of the original models are replicated elsewhere, the models would need to be recalibrated before the risk estimates could be clinically useful.[25] After recalibration in our training cohort, all models demonstrated improved calibration but logCASUS was slightly better calibrated than SOFA and RACE.

This study represents the first external validation of the logCASUS and RACE models and the first validation of the SOFA score in UK cardiac surgery. We utilised contemporary data from a tertiary cardiac centre with excellent clinical results and undertook a comprehensive assessment of model performance. As with any clinical study, there were missing data but the proportion in this study was low. The variables with the most missing data were PAR (required for logCASUS calculation only) and serum bilirubin. A clinically robust approach to handling missing data was adopted.

A potential limitation of the study is that it is based on data from a single centre and includes relatively few outcomes. The division of our dataset for development and evaluation of locally recalibrated versions of the models exacerbated this problem and so we limited the evaluation of the recalibrated models to the first five postoperative days. When the number of outcomes is low, validation results and performance statistics need to be interpreted with caution. Potential options to increase the sample size would be to expand the number of participating centres or increase the timeframe for data collection. The first option is not feasible at present due to a lack of UK centres currently collecting the necessary data in a suitable format. Although the second option is feasible, this would introduce temporal issues related to changes in practice and performance over time that could affect the results.[25]

The SOFA score was the first of the validated models to be developed and was designed based on expert consensus in 1996 to assess the progress of patients suffering from sepsis.[15] It has since been validated as a prediction tool for adverse

outcomes in general ICU patients[26] and also specifically in patients who have undergone cardiac surgery.[7, 10] It grades the function of six organ systems using a five point scale for each with totals ranging between zero and 24 (Table 2).

The logistic logCASUS[18] and RACE[19] scores were developed by the same team as the additive CASUS score[4] using data from a single centre. The RACE score was designed as a user friendly version of logCASUS. Unlike SOFA, these scores were developed exclusively for patients who have undergone cardiac surgery. They both include use of mechanical circulatory support and intra-aortic balloon pump counterpulsation which are more common in cardiac surgery patients than the general ICU population. Both scores grade neurological status using a scale which reduces the impact of the low conscious level expected immediately after cardiac surgery and adjust the predicted risk based on the time since surgery.

During the recalibration process we were able to identify which variables within each model were significantly associated with ICU mortality. We found that variables which can be controlled by physicians such as MAP and PAR were not significantly associated with outcome. Conversely interventions which aim may affect those parameters such as use of mechanical ventilation, renal replacement therapy, or mechanical circulatory support were shown to have significant and relatively large effects on risk. Serum creatinine and lactate concentrations and the platelet count were the most significant of the blood analyses assessed. (Appendices B-D)

Despite generally superior discriminatory ability compared to preoperative models,[27] none of the validated models has been widely adopted in clinical practice. Possible reasons for this include problems with the ability to easily calculate the scores, a perceived lack of clinical utility or validity and inadequate external validation studies. Clinical utility of the scores is limited by the fact that they are inherently retrospective as

they rely on the worst value obtained over a 24-hour period. Importantly, the cardiac surgery-specific scores are designed to be used only during a patient's first admission to CICU. In addition, in these models the first postoperative day is the reference for the "ICU day" variable, i.e. the beta-coefficient for postoperative day one is zero. The models do not provide an "ICU day" coefficient for the operative day and consequently cannot produce risk estimates until the first postoperative day. As a result of these limitations in model design, the logCASUS and RACE risk scores could not be calculated for 29 patients for whom the first ICU episode finished on the operative day. This is clinically relevant because for 24 of these patients, the short initial admission was due to reoperation for bleeding and the ICU mortality rate in this group was 17%. Although this did not significantly change the overall ICU mortality rate (2.0% versus 1.8%) the inability to assess risk in these patients is a limitation of the models.

Despite limitations, the models studied could potentially be utilised for three main purposes. All models discriminate well between patients at high and low risk of mortality therefore clinicians could use the scores to identify patients with the highest risk amongst those present on CICU and to target resources including staff allocations accordingly. Secondly, if validated in multicentre-studies the models could be utilised for the risk-adjustment of CICU benchmarking data in a similar way to that in which preoperative models have been used to risk-adjust surgical outcome data. [28] Utilising models that include postoperative variables to generate risk predictions allows risk estimates to be modified by the occurrence intra- and early postoperative events. Such estimates would be better suited for the assessment of CICU performance. If the models were to be used for this purpose risk predictions should be made early in the CICU stay because scores calculated later are likely to be influenced by the quality of care already received on the CICU.

Lastly, if the models are appropriately calibrated locally, daily scores could be used to assess patient progress, assist clinical decision making and to inform discussions with patients and their relatives by providing the most up to date assessment of patient risk possible. Although predictions will never be completely accurate for individual patients, they may be used as a guide.

In conclusion, all three models showed good discrimination when used during the first postoperative week after cardiac surgery. In their original forms the cardiac surgery specific models were poorly calibrated, particularly in patients with the highest risk, but all three models could be successfully recalibrated using local data. Further research into optimising postoperative models to maximise their clinical utility is required if they are going to be widely adopted.

Funding Statement

Work performed producing this manuscript was funded by British Heart Foundation grant number PG/16/80/32411.

References

[1] Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardiothorac Surg 1999;16:9-13.

[2] Roques F, Michel P, Goldstone AR, Nashef SA. The logistic EuroSCORE. Eur Heart J 2003;24:881-2

[3] Nashef SA, Roques F, Sharples LD et al. EuroSCORE II. Eur J Cardiothorac Surg 2012;41:734-744; discussion 744-735.

[4] Hekmat K, Kroener A, Stuetzer H, et al. Daily assessment of organ dysfunction and survival in intensive care unit cardiac surgical patients. Ann Thorac Surg 2005;79:1555-1562.

[5] Howitt SH, Grant SW, Riding DM, Malagon I, McCollum CN. (in press) Risk models that utilise postoperative patient monitoring data to predict outcomes in adult cardiac surgery; a systematic review. J Cardiothorac Vasc Anesth.

[6] Hekmat K, Doerr F, Kroener A et al. Prediction of mortality in intensive care unit cardiac surgical patients. Eur J Cardiothorac Surg 2010;38:104-109.

[7] Doerr F, Badreldin AM, Heldwein MB et al. A comparative study of four intensive care outcome prediction models in cardiac surgery patients. J Cardiothorac Surg.2011;6. DOI:10.1186/1749-8090-6-21

[8] Ceriani R, Mazzoni M, Bortone F et al. Application of the sequential organ failure assessment score to cardiac surgical patients. Chest 2003;123:1229-1239.

[9] Patila T, Kukkonen S, Vento A, Pettila V, Suojaranta-Ylinen R. Relation of the Sequential Organ Failure Assessment Score to Morbidity and Mortality After Cardiac Surgery. Ann Thorac Surg 2006;82:2072-2078.

[10] Gomes RV, Tura B, Mendonca HT et al. A first postoperative day predictive score of mortality for cardiac surgery. Ann Thorac Cardiovasc Surg 2007;13:159-164.

[11] Badreldin A, Elsobky S, Lehmann T, Brehm BB, Doenst T, Hekmat K. Daily-mean-SOFA, a new derivative to increase accuracy of mortality prediction in cardiac surgical intensive care units. Thorac Cardiovasc Surg 2012;60:43-50.

[12] Doerr F, Badreldin AM, Can F, Bayer O, Wahlers T, Hekmat K. SAPS 3 is not superior to SAPS 2 in cardiac surgery patients. Scand Cardiovasc J 2014;48:111-119.

[13] Ariyaratnam P, Loubani M, Biddulph J et al. Validation of the intensive care national audit and research centre scoring system in a UK adult cardiac surgery population. J Cardiothorac Vasc Anesth 2015;29:565-569.

[14] Heldwein MB, Badreldin AM, Doerr F et al. Logistic Organ Dysfunction Score (LODS): a reliable postoperative risk management score also in cardiac surgical patients? J Cardiothorac Surg.2011;6. DOI:10.1186/1749-8090-6-110

[15] Vincent JL, Moreno R, Takala J et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on

Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med 1996;22:707-710.

[16] Badreldin AM, Doerr F, Ismail MM, et al. Comparison between Sequential Organ Failure Assessment score (SOFA) and Cardiac Surgery Score (CASUS) for mortality prediction after cardiac surgery. Thorac Cardiovasc Surg 2012;60(1):35-42.

[17] Exarchopoulos T, Charitidou E, Dedeilias P, Charitos C, Routsi C. Scoring Systems for Outcome Prediction in a Cardiac Surgical Intensive Care Unit: A Comparative Study. Am J Crit Care 2015;24:327-334.

[18] Doerr F, Badreldin AM, Bender EM et al. Outcome prediction in cardiac surgery: the first logistic scoring model for cardiac surgical intensive care patients. Minerva Anestesiol 2012;78:879-886.

[19] Badreldin AM, Doerr F, Bender EM et al. Rapid clinical evaluation: an early warning cardiac surgical scoring system for hand-held digital devices. Eur J Cardiothorac Surg 2013;44:992-997.

[20] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing 2015.

[21] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.

[22] Hosmer DW, Lemeshow S. Applied logistic regression(2nd Edition): Wiley,2000;147-155.

[23] Bridgewater B, Grant SW, Hickey GL, et al. National Adult Cardiac Surgery Audit Report: National Institute for Cardiovascular Outcomes Research; 2012 2012.

[24] Hickey GL, Grant SW, Murphy GJ et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. Eur J Cardiothorac Surg 2013;43:1146-1152.

[25] Grant SW, Grayson AD, Jackson M et al. Does the choice of risk-adjustment model influence the outcome of surgeon-specific mortality analysis? A retrospective analysis of 14637 patients under 31 surgeons. Heart 2008;94:1044-1049.

[26] Ferreira FL, Bota DP, Bross A, Melot C, Vincent JL. Serial evaluation of the SOFA score to predict outcome in critically ill patients. JAMA 2001;286:1754-1758.

[27] Grant SW, Hickey GL, Dimarakis I et al. How does EuroSCORE II perform in UK cardiac surgery; an analysis of 23 740 patients from the Society for Cardiothoracic Surgery in Great Britain and Ireland National Database. Heart 2012;98:1568-1572.

[28] Grant SW, Hickey GL, Cosgriff R et al. Creating transparency in UK adult cardiac surgery data. Heart 2013;99:1067-1068.

Figure legend

Figure 1 (a) Receiver Operating Characteristic (ROC) curves for the validated models on the first postoperative day. The dashed line represents model performance no better than random chance. (b) calibration plots for the original logCASUS and RACE models and recalibrated logCASUS, RACE and SOFA models on the first postoperative day. The dashed line represents the line of perfect calibration.